Nelralec/Bhasha Sanchar Working Paper 2
Categorisation for automated morphosyntactic analysis of Nepali: introducing the Nelralec Tagset (NT-01)

by Andrew Hardie, Ram Lohani, Bhim Regmi and Yogendra Yadava

July 2005

1. Introduction

This working paper presents a summary of the Nelralec Tagset, a categorisation system for the manual and automated analysis of morphosyntactic units in Nepali[1].

This tagset has been compiled with reference to widely published grammars of Nepali. These grammars in the main describe standard Nepali. The tagset may not be sufficient for other dialects of Nepali[2].

In sections 2 to 4, the most important issues underlying the construction of the categories in the tagset are discussed: the handling of case affixes/postposition, the handling of gender, and the model of Nepali verb inflection underlying the tagset categories. In section 5, the tagset itself is tabulated with examples.

2. Tokenisation and postpositions

The first step in both manual and automated tagging is the division of the text into *tokens*. Tokens are often described as "words" but in this context it may be better to think of them as "appropriately-sized units for morphosyntactic analysis"; such a unit may be larger than, or smaller than, an orthographic word (i.e. a string of letters bounded by white space or punctuation).

For the purposes of Nepali tagging, there is one major instance where words must be split apart to create appropriately-sized units for analysis: a group of morphemes which are variously considered to be clitics, case affixes, or postpositions.

While in many languages case, gender and number are best analysed together, in Nepali case and number are indicated by elements similar to (and tagged similarly to) postpositions, whereas gender is marked by an inflected affix. Thus, distinct considerations apply to gender and to case-number, (see below on gender).

In Nepali, there are a large number of forms that are considered to be *either* postpositions *or* case-suffixes. There is no 100% agreement on what is or is not a suffix. Some Nepali grammatical traditions consider the following forms to be suffixes: *haruu*, *ko/kii/kaa*, *le*, *laaii* (whose meanings are, respectively, plural/collective; genitive; ergative/instrumental; and accusative/dative), and

[1] It should be noted that this published document represents an abbreviation of the full tagset description used for manual and automatic tagging on the Nelralec project.
[2] When an example in the tagset definition is enclosed in [square brackets], that example has been reported to be limited to particular dialects.

consider all other postposition/suffix forms to be postpositions. However, in terms of part-of-speech tagging it is difficult to draw a consistent distinction between *haruu*, *ko/kii/kaa*, *le*, *laaii* and other postpositions such as *maa*, *baaTa, sanga, dekhi*, etc., because the combinations in which these morphemes can occur are too various to allow them to be classified using variations in the tag for the form that they are attached to[3].

For this reason, *all* these forms are tagged separately to the noun, adjective, pronoun or other form to which they are attached. This means that postpositions are necessarily tokenised separately to the forms to which they are attached, and tagged separately.

3. Gender on nouns and adjectives

Some Nepali categories mark gender by means of a three-way *o/ii/aa* distinction, where *o* is masculine, *ii* is feminine, and *aa* is "other" (it can indicate masculine plural, or feminine plural, or oblique case marking motivated by a following postposition, or honorific marking). There is also one postposition, *ko/kii/kaa*, which is marked for gender in this way.

However, this gender marking is not found on all members of the gendered categories. Many words are "unmarked" – that is, they have a single invariant form regardless of gender. For example, adjectives are a gendered category, speaking generally: they are marked for gender agreement with the noun they modify. But many, many adjectives are actually *unmarked* – that is, there is nothing in their morphology to indicate their gender.

Nepali nouns have *natural* gender rather than grammatical gender. Animate females are feminine, all other nouns are masculine. In most cases gender is not marked on nouns in the way that it is marked on some adjectives. For a minority of nouns, there are pairs of masculine and feminine nouns related through the *–o/ii* distinction, for instance *keTo* "boy" / *keTii* "girl". But there are also numerous feminine nouns that end in *ii* without there being a masculine equivalent in *o* (e.g. *aaimaaii*, "woman").

Thus the *o/ii/aa* distinction is ignored on nouns for the purpose of this system of POS tagging, but *not* on pronouns, adjectives, non-finite verbs, etc., where the distinction is motivated by agreement and is thus clearly inflectional rather than lexical-derivational, as is the case for nouns.
Gender is not included in the tags for nouns because it is not an inflectional category in the same way that it is for adjectives.

4. Modelling Nepali verb inflections

Nepali has a vast number of verb inflections. This is because compounding is an extremely productive process in the verb system of Nepali. Different combinations of various non-finite forms of the main verb and various inflected forms of secondary verbs creates a very large number of

[3] The following convention is used in this document when exemplifying tokens that result from the splitting-up of another token: when an example begins or ends in the symbol #, this indicates that the form being exemplified is a form that has been separated – i.e. it is either a morpheme which has been separated, or a form from which a morpheme has been separated.

tense-aspect-mood combinations[4]. If every distinction made in the verb system were to be indicated by a separate tag, then the tags for verbs would become entirely unmanageable. Thousands of tags would be required for the verbs alone. The simplified classification model of the Nepali verb that this tagset uses is described in this section.

To avoid the problem of compound verbs creating an unmanageable number of categories, the tagging system described here uses this rule for verbs:

Every compound verb is tagged according to the last verb in the compound.

That means that only the last identifiable verb in a compound verb is taken into account when deciding the tag. If there is only one identifiable verb, then the whole thing is taken into account.

As a further simplifying measure, certain aspects of the verb morphology system are ignored. These are passive, causative, and negative. This means that:

- Any passive verb is tagged the same as the corresponding active verb
  - E.g. tag *garinu* the same as *garnu*
- Any causative verb is tagged the same as the corresponding non-causative verb
  - E.g. tag *garaau~cha* the same as *garcha*
- Any negative verb is tagged the same as the corresponding positive verb
  - E.g. tag *garnechaina* the same as *garnecha*

No distinction is made in the tagging between auxiliary verbs and main verbs.

The tagset makes, for convenience of reference only, a distinction between finite (person-marked) verb forms, and non-finite (non-person-marked) verb forms. Many of the non-finite forms occur embedded at the start of a longer verb (e.g. **gardai**thyo, **garnu**huncha). However, in accordance with the general rule, they are **not** tagged separately in this case. This means that non-finite tags are **only** used *if the non-finite form is written as a separate word*, or *if the non-finite form is at the end of the longer verb*.

The non-finite forms distinguished by tagset are as follows:

- The infinitive
- A group of participles[5]:
  - the *e(ko)*-participle, sometimes called the *perfect participle*
  - the *d*-participle (*do/dii/daa/dai*), which is used for three functions:
    - the converb/participle function (also known as the *conjunctive participle*, the *progressive participle*, and the *simultaneous converb*)
    - the modifier function
    - as an element of compound verbs (where it does not receive a separate tag)

[4] The term "compound verb" is used in a wide sense here, to refer to *any verb that consists of two or more elements that are identifiable as belonging to separate independent verb lexemes*. So, for instance, *garidiyo* would be a straightforward example of a compound verb; however, for the purposes of this description, words such as *garthyo*, *garcha*, *garirahyo*, and *garnecha* will be considered to be compound verbs as well. Compounds consisting of noun+verb or adjective+verb can be tagged using the same rule: only consider the second element, i.e. the last verb.
[5] They should probably not all be considered "participles" in the most precise grammatical sense, Some are "converbs" and some may be more precisely analysed as infinitival.

    - the *ne*-participle (the *imperfect participle* or *infinitival participle*)
    - the sequential converbs, also called *absolutive participles*, of which there are three, which all receive the same tag:
        - the *era*-participle
        - the *ii*-participle
        - the *iikana*-participle
- Two other forms:
    - The *e*-form (often referred to as *subjunctive* or *conditional*)
    - the *i*-form, sometimes referred to as the *passive root*, which is the form in which a non-final verb in a compound appears.
- Three command forms (also called the *imperative*)

The finite forms are modelled as follows:

There are six finite forms for each tense/mood/aspect combination, which are indicated in the third person singular by the following typical endings: *-yo, -thyo, -echa, -cha, -necha, -laa*. Many of these tenses follow similar patterns of inflection. An additional finite form, the *optative*, is given separate tags.

The following distinctions operate on finite verbs:

- Person: first, second, third
- Gender: masculine (default), feminine
- Number: singular, plural
- Honorific level: non-honorific, medial-honorific[6]

Therefore, in theory there should be 3 x 2 x 2 x 2 = 24 tags. However, not all possible combinations of features have separate forms in Nepali. There is, for instance, no specific form in any tense-mood for the second-person non-honorific feminine plural. There is just a single second-person plural form.

In general, only second and third person singular verbs are marked for gender and honorific level. Plural verb forms are not marked for gender or for honorific level. First person verbs are not marked for gender or for honorific level. Similarly, at the medial-honorific level, there is no distinction between singular and plural, except for feminine verbs (which are listed separately below). This means that if we take only gender, number and honorific level into account, and if we treat alike those categories which are merged together, only ten tags are necessary.

In many cases gender is not marked on verbs. Even where it is marked, it is reported that the "masculine" verbs may be used in some varieties of Nepali with feminine subjects. For this reason, the tagset considers "masculine" to be the default option. So the six "masculine" tags listed in the table below, are in fact not defined as masculine: they are simply the usual, default forms. The four distinct feminine forms– the second person and third person singular, non-honorific and medial-honorific – are defined in contrast to the default: the tags for them are created by adding F to the tag.

There are separate tags for optative verbs, as these verbs behave differently in many ways to the other finite verbs (e.g. by taking a prefix to indicate the negative, rather than a suffix). However, the

[6] The higher honorific levels (high, royal) are conveyed through compound verbs.

person-number-honorific categories in the optative paradigm are directly parallel to those of the general finite paradigm. There are no feminine forms in the optative, and thus no tags for them.

5. The tagset

The following table gives a summary of the tagset. It contains 112 tags.

Some of the categories in this tagset are open, and some are closed. In terms of the tagset, this simply means whether or not an exhaustive listing of the word-forms that can take that tag is possible. A *closed category* is one where it is possible to list the contents exhaustively; an *open category* is one where it is not. The tags RR and UU, and any tags beginning with N, J, F (except FZ), or V, indicate **open categories**. The tags FZ, RD, RJ and RK, and any tags beginning with C, I, P, D, M, or Y, indicate **closed categories**. Examples given in *italics* represent a closed category. Not all closed categories are listed exhaustively in the table below.

Category definition	**Examples (Latin)**	**Examples (Devanagari)**	**Tag**
Common noun	keTo, keTaa, kalam	केटो, केटा कलम	NN
Proper noun	raam	राम	NP
Masculine adjective	moTo, raamro	मोटो, राम्रो	JM
Feminine adjective	moTii, raamrii	मोटी, राम्री	JF
Other-agreement adjective	moTaa, raamraa	मोटा, राम्रा	JO
Unmarked adjective	saphaa, dhanii, asal	सफा, धनी, असल	JX
Sanskrit-derived comparative or superlative adjective	uccatar, uccatam	उच्चतर, उच्चतम	JT
First person pronoun	*ma, haamii, mai#*	म, हामी, मै#	PMX
First person possessive pronoun with masculine agreement	*mero, haamro*	मेरो, हाम्रो	PMXKM
First person possessive pronoun with feminine agreement	*merii, haamrii*	मेरी, हाम्री	PMXKF
First person possessive pronoun with other agreement	*meraa, haamraa*	मेरा, हाम्रा	PMXKO
Non-honorific second person pronoun	*ta~, tai#*	तँ, तैँ#	PTN
Non-honorific second person possessive pronoun with	*tero*	तेरो	PTNKM

masculine agreement			
Non-honorific second person possessive pronoun with feminine agreement	*terii*	तेरी	PTNKF
Non-honorific second person possessive pronoun with other agreement	*teraa*	तेरा	PTNKO
Medial-honorific second person pronoun	*timii*	तिमी	PTM
Medial-honorific second person possessive pronoun with masculine agreement	*timro*	तिम्रो	PTMKM
Medial-honorific second person possessive pronoun with feminine agreement	*timrii*	तिम्री	PTMKF
Medial-honorific second person possessive pronoun with other agreement	*timraa*	तिम्रा	PTMKO
High-honorific second person pronoun	*tapaai~, hajur*	तपाईँ, हजुर	PTH
High-honorific unspecified-person pronoun	*yahaa~, wahaa~*[7]	यहाँ, वहाँ	PXH
Royal-honorific unspecified-person pronoun	*sarkaar, mausuph*	सरकार, मौसुफ	PXR
Reflexive pronoun	*aaphuu*	आफू	PRF
Possessive reflexive pronoun with masculine agreement	*aaphno*	आफ्नो	PRFKM
Possessive reflexive pronoun with feminine agreement	*aaphnii*	आफ्नी	PRFKF
Possessive reflexive pronoun with other agreement	*aaphnaa*	आफ्ना	PRFKO
Masculine demonstrative determiner	*yasto, yatro*	यस्तो, यत्रो,	DDM
Feminine demonstrative determiner	*yastii, yatrii*	यस्ती, यत्री	DDF
Other-agreement demonstrative determiner	*yastaa, yatraa*	यस्ता, यत्रा	DDO
Unmarked demonstrative determiner	*yo, yas#, yi, yin#, yinii, yati, yatti*	यो, यस#,यी, यिन#,	DDX

[7] There is an alternative form for *wahaa~*: *uhaa~*, उहाँ. This form would also take PXH.

		यिनी, यति, यत्ति	
Masculine interrogative determiner	*kasto, katro*	कस्तो, कत्रो	DKM
Feminine interrogative determiner	*kastii, katrii*	कस्ती, कत्री	DKF
Other-agreement interrogative determiner	*kastaa, katraa*	कस्ता, कत्रा	DKO
Unmarked interrogative determiner	*ko, kas#, ke, kun, kati*	को, कस#, के, कुन, कति	DKX
Masculine relative determiner	*jasto, jatro*	जस्तो, जत्रो,	DJM
Feminine relative determiner	*jastii, jatrii*	जस्ती, जत्री	DJF
Other-agreement relative determiner	*jastaa, jatraa*	जस्ता, जत्रा	DJO
Unmarked relative determiner	*jo, jas#, je, jati, josukai*	जो, जस#, जे, जति, जोसुकै	DJX
Masculine general determiner-pronoun	arko	अर्को	DGM
Feminine general determiner-pronoun	arkii	अर्की	DGF
Other-agreement general determiner-pronoun	arkaa	अर्का	DGO
Unmarked general determiner-pronoun	aruu	अरू	DGX
Infinitive verb	garnu, garna, garnaa, nagarnu, nagarna, nagarnaa	गर्नु, गर्न, गर्ना, नगर्नु, नगर्नु, नगर्ना	VI
Masculine d-participle verb	gardo, nagardo	गर्दो, नगर्दो	VDM
Feminine d-participle verb	gardii, nagardii	गर्दी, नगर्दी	VDF
Other-agreement d-participle verb	gardaa, nagardaa	गर्दा, नगर्दा	VDO
Unmarked d-participle verb	gardai, nagardai	गर्दै, नगर्दै	VDX
e(ko)-participle verb	gae (as in gae saal), gare (as in garejati or gareko)	गरे	VE
ne-participle verb::	garne, nagarne	गर्ने, नगर्ने	VN
Sequential participle-converb	garera, gariikana, garii, nagarera,	गरेर, गरीकन, गरी,	VQ

	nagariikana, nagarii	नगरेर, नगरीकन, नगरी	
Command-form verb, non-honorific	gar, jaa	गर्, जा	VCN
Command-form verb, mid-honorific	gara, jaau, jaao	गर, जाऊ, जाओ	VCM
Command-form verb, high-honorific	garnos, jaanos	गर्नोस्, जानोस्	VCH
Subjunctive / conditional e-form verb	gare, nagare	गरे, नगरे	VS
i-form verb	gari	गरि	VR
First person singular verb	gare~, garthe~, garina~, chu, hu~, garnechu	गरँ, गर्थँ ,गरिनँ, छु, हुँ, गर्नेछु	VVMX1
First person plural verb	garyau~, garthyau~, garenau~, chau~, hau~, garnechau~	गर्यौँ, गथ्यौँ, गरेनौँ, छौँ, हौँ, गर्नेछौँ	VVMX2
Second person non-honorific singular verb	garis, garthis, garinas, chas, hos, garnechas	गरिस्, गर्थिस्, गरिनस्, छस्, होस्, गर्नेछस्	VVTN1
Second person plural (or medial-honorific singular) verb	garyau, garthyau, garenau, chau, hau, garnechau	गर्यौ, गथ्यौ, गरेनौ, छौ, हौ, गर्नेछौ	VVTX2
Third person non-honorific singular verb	garyo, garthyo, garena, cha, ho, garnecha	गर्यो, गथ्यो, गरेन, छ, हो, गर्नेछ	VVYN1
Third person plural (or medial-honorific singular) verb	gare, garthe, garenan, chan, hun, garnechan	गरे, गर्थे, गरेनन्, छन्, हुन्, गर्नेछन्	VVYX2
Feminine second person non-honorific singular verb	garlis, ches, garthis	गर्लिस्, छेस्, गर्थिस्	VVTN1F
Feminine second person non-honorific singular verb	garthyau, chyau	गथ्यौ, छ्यौ	VVTM1F
Feminine third person medial-honorific singular verb	garina, garii, che, garthii	गरिन, गरी, छे, गर्थी	VVYN1F
Feminine third person medial-honorific singular verb	garin, garthin, garinan, chin	गरिन्, गर्थिन्, गरिनन्, छिन्	VVYM1F
First person singular optative verb	jaau~, garu~	जाऊँ, गरूँ	VOMX1
First person plural optative verb	jaaau~, garau~	जाऔँ, गरौँ	VOMX2

Second person non-honorific singular optative verb	gaes, gares	गएस्, गरेस्	VOTN1
Second person plural (or medial-honorific singular) optative verb	gae, gare	गए, गरे	VOTX2
Third person non-honorific singular optative verb	jaaos, garos	जाओस्, गरोस्	VOYN1
Third person plural (or medial-honorific singular) optative verb	jaauun, garuun	जाऊन्, गरून्	VOYX2
Adverb	raamrarii, ekdam, chiTo	राम्ररी, एकदम, छिटो	RR
Demonstrative adverb	yataa, utaa, tyataa; ahile, ahilyai, yahii~, yasarii, aba	यता, उता, त्यता, अहिले, अहिल्यै, यहीं, यसरी, अब	RD
Interrogative adverb	kataa, kahaa~, kahile, kasarii	कता, कहाँ, कहिले	RK
Relative adverb	jataa, jahaa~, jahile, jasarii, jaba	जता, जहाँ, जहिले	RJ
Postposition	agaaDi, pachaaDi, baaTa, dwaaraa, maa, maathi, saath, puurvak, tira, tarpha, vasa, sanga, binaa	अगाडि, पछाडि, बाट, द्वारा, मा, माथि	II
Plural-collective postposition	*haruu*	हरू	IH
Ergative-instrumental postposition	*le*	ले	IE
Accusative-dative postposition	*laaii*	लाई	IA
Masculine genitive postposition	*ko*	को	IKM
Feminine genitive postposition	*kii*	की	IKF
Other-agreement genitive postposition	*kaa*	का	IKO
Cardinal number	*ek, eu#, yau#, dui, tin, caar, paa~c*	एक, दुई, तीन, चार, पाँच	MM
Masculine ordinal number	*pahilo, dosro, tesro, cautho*	पहिलो, दोस्रो, तेस्री, चौथो	MOM
Feminine ordinal number	*pahilii, dosrii, tesrii, cauthii*	पहिली, दोस्री, तेस्री,	MOF

		चौथी	
Other-agreement ordinal number	*pahilaa, dosraa, tesraa, cauthaa*	पहिला, दोस्रा, तेस्रा, चौथा	MOO
Unmarked ordinal number	*paa~cau~*	पाँचौँ	MOX
Masculine numeral classifier	*#To* [as in euTo]	#टो	MLM
Feminine numeral classifier	*(#)waTii, #Tii, #oTii, #auTii*	(#)वटी, #टी, #ओटी	MLF
Other-agreement numeral classifier	*(#)waTaa, #Taa, #oTaa, #auTaa*	(#)वटा, #टा, #ओटा, #औटा	MLO
Unmarked numeral classifier	*(#)janaa*	(#)जना	MLX
Coordinating conjunction	ra, tathaa	र, तथा	CC
Subordinating conjunction appearing **after** the clause it subordinates	bhanne, bhanii, bhanera	भने, भनी, भनेर	CSA
Subordinating conjunction appearing **before** the clause it subordinates	ki, yadi, yaddyapi, kinaki	कि, यदि, यद्यपि, किनकि	CSB
Particle	nai, caahi~, pani, hai, ra, re, kyaare, ho, khai	नै, चाहिँ, पनि, है, र, रे, क्यारे, हो, खै	TT
Question marker	*ke*	के	QQ
Interjection	oho, aahaa, hare	ओहो, आहा, हरे	UU
Possessive reflexive pronoun without agreement	*merai*	मेरै	PMXKX
Non-honorific second person possessive pronoun without agreement	*terai*	तेरै	PTNKX
Medial-honorific second person possessive pronoun without agreement	*timrai*	तिम्रै	PTMKX
Possessive reflexive pronoun without agreement	*aaphnai*	आफ्नै	PRFKX
Unmarked genitive postposition	*kai*	कै	IKX
Sentence-final punctuation	? ! . । ॥		YF

Sentence-medial punctuation	, ; : :- / –		YM
Quotation marks	' "		YQ
Brackets	( ) { } [ ]		YB
Foreign word in Devanagari			FF
Foreign word, not in Devanagari			FS
Abbreviation	M.P.P.	म. पु. पु.	FB
Mathematical formula (and similar)	$e=mc^2$		FO
Letter of the alphabet			FZ
Unclassifiable			FU
Null tag: an element of the text which does not need a tag	<p>		NULL